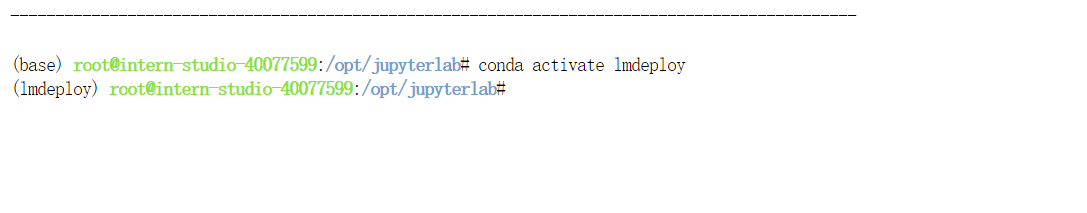
基础作业（结营必做）

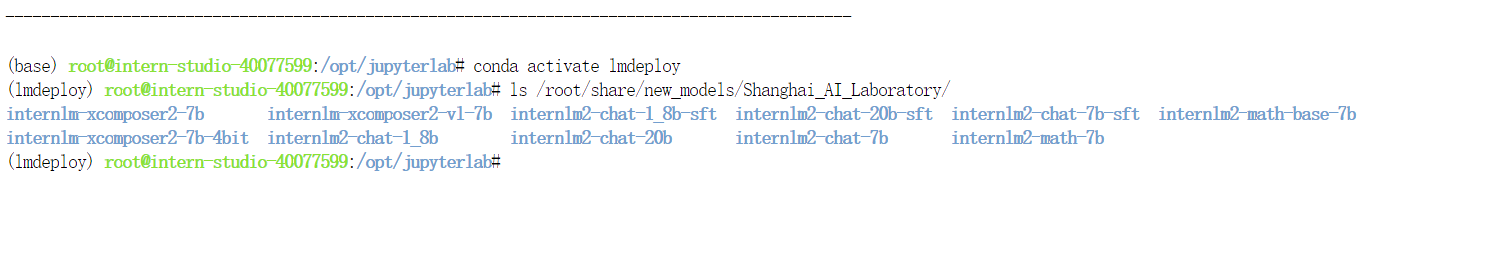
完成以下任务，并将实现过程记录截图：

配置lmdeploy运行环境

conda activate lmdeploy



ls /root/share/new\_models/Shanghai\_AI\_Laboratory/



以命令行方式与模型对话

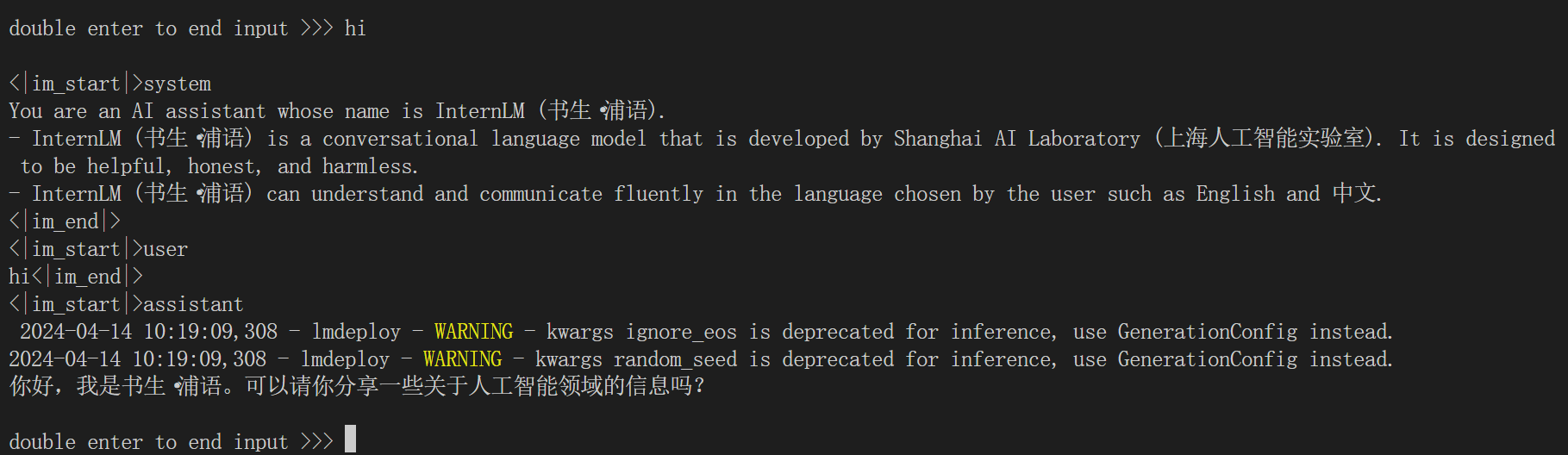
lmdeploy chat /root/internlm2-chat-1\_8b

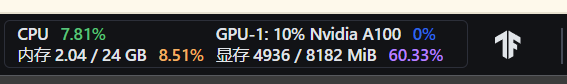
进阶作业

完成以下任务，并将实现过程记录截图：

1.设置KV Cache最大占用比例为0.4，开启W4A16量化，以命令行方式与模型对话。（优秀学员必做）

lmdeploy chat /root/internlm2-chat-1\_8b-4bit --model-format awq --cache-max-entry-count 0.4





2.以API Server方式启动 lmdeploy，开启 W4A16量化，调整KV Cache的占用比例为0.4，分别使用命令行客户端与Gradio网页客户端与模型对话。（优秀学员）

命令行客户端

lmdeploy serve api\_server \

/root/internlm2-chat-1\_8b-4bit \

--model-format awq \

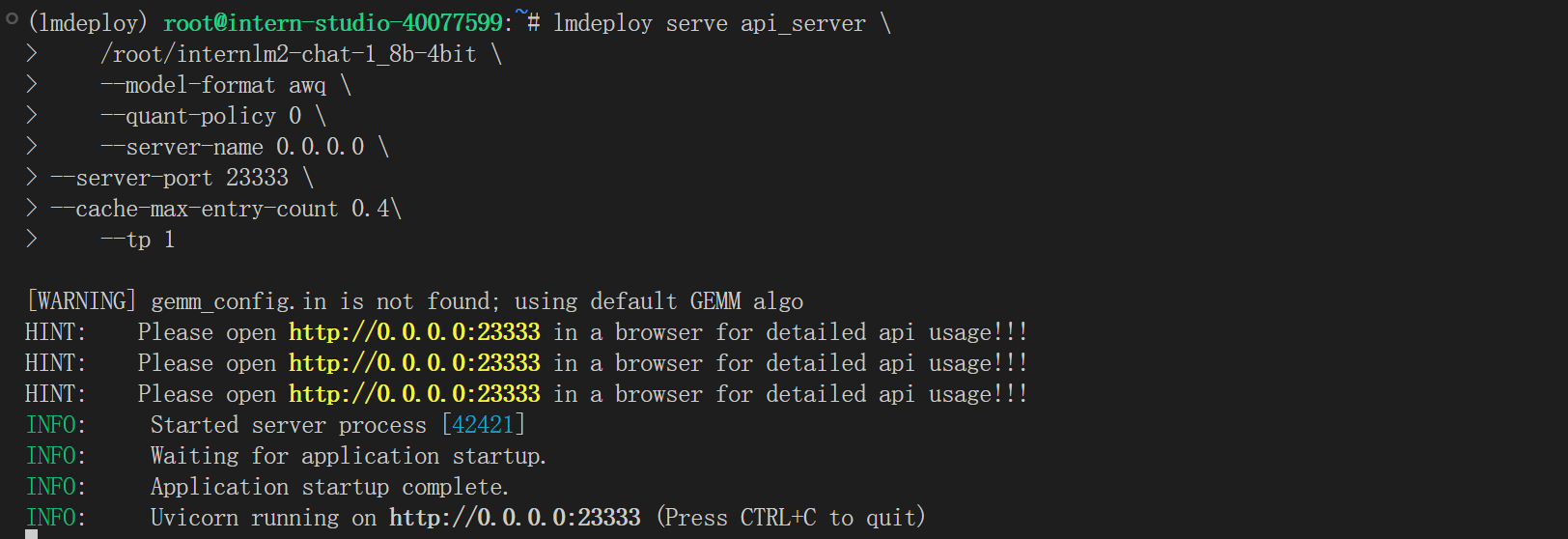
--quant-policy 0 \

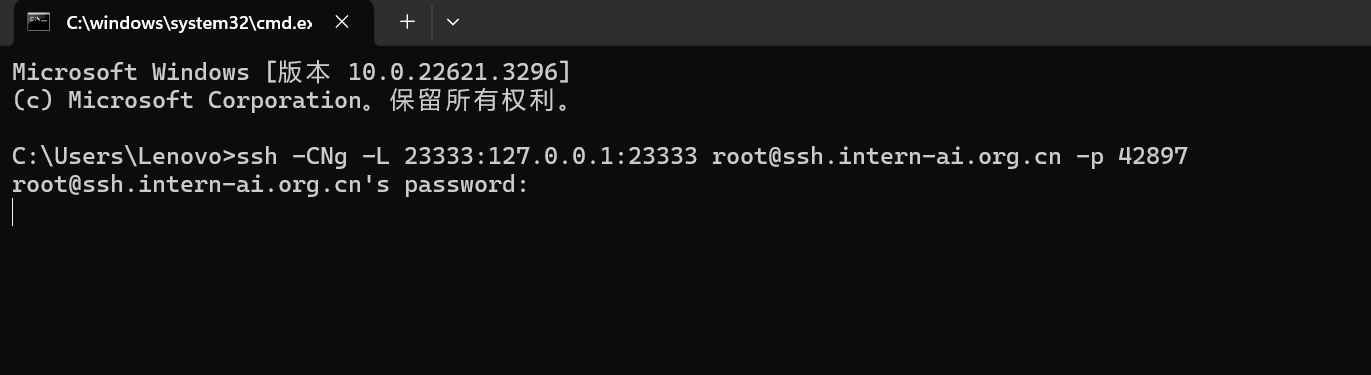
--server-name 0.0.0.0 \

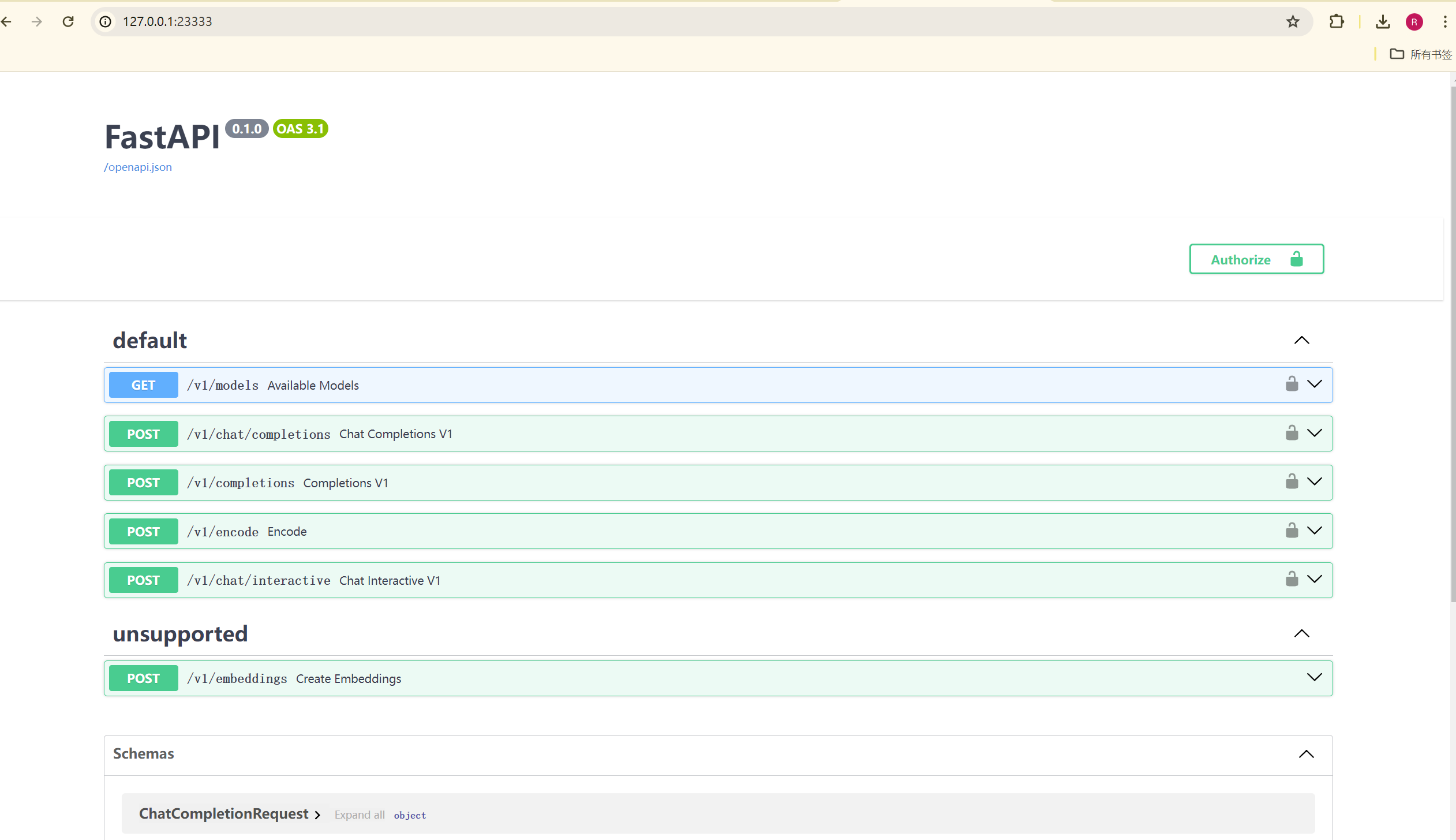
--server-port 23333 \

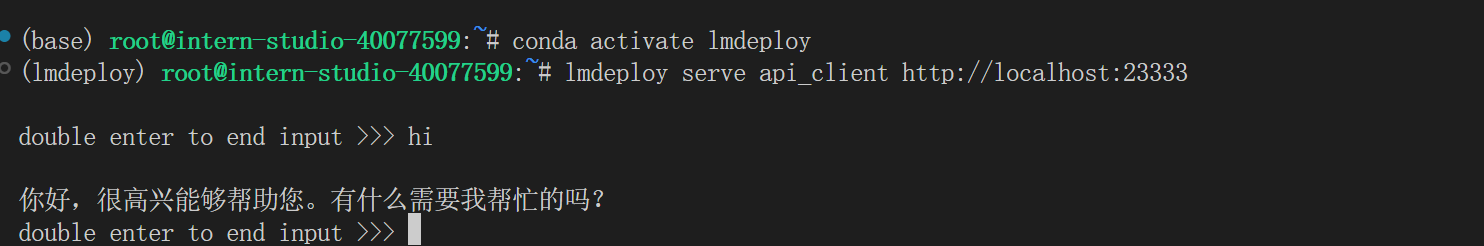
--cache-max-entry-count 0.4\

--tp 1







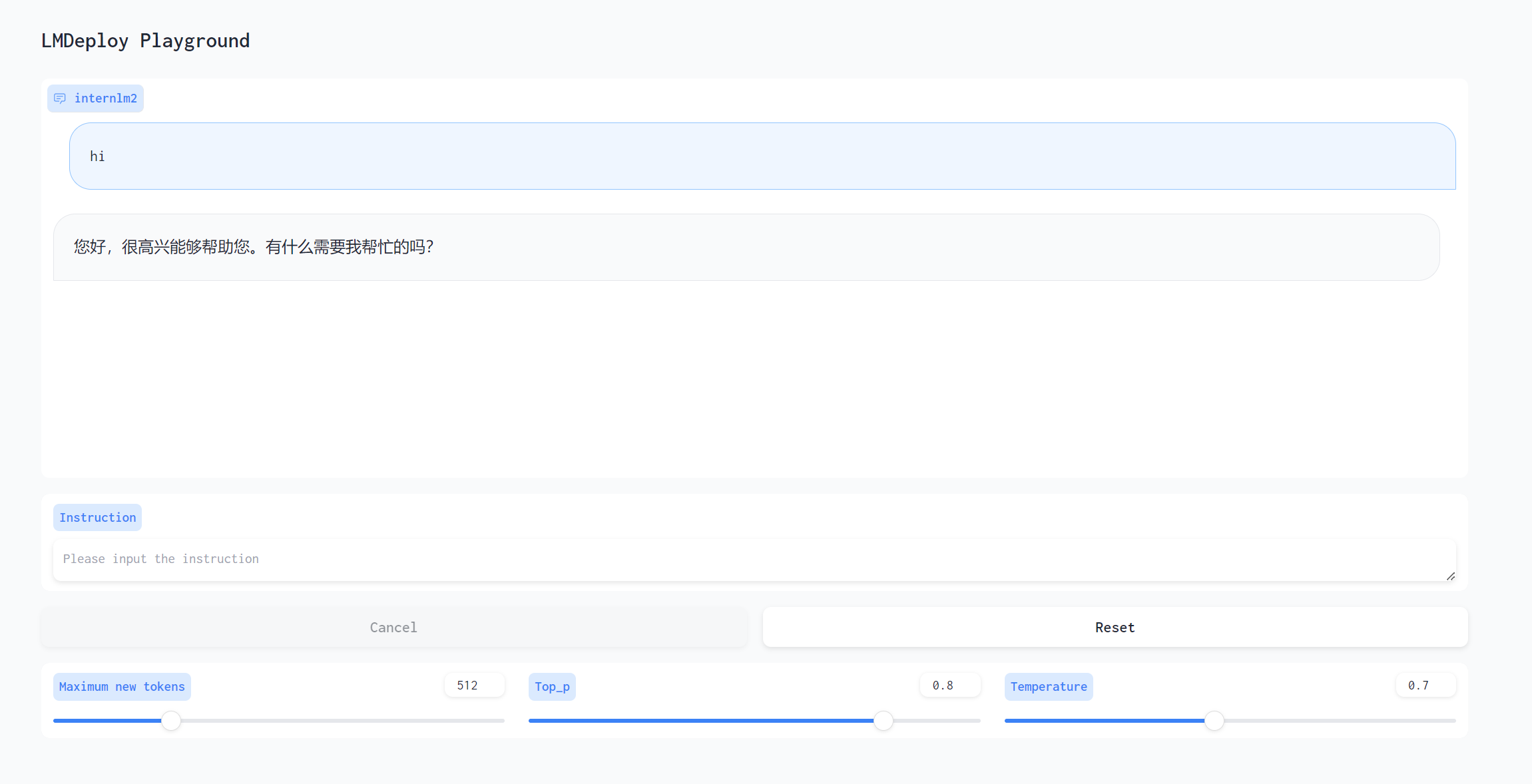


网页客户端连接API服务器

lmdeploy serve gradio http://localhost:23333 \

--server-name 0.0.0.0 \

--server-port 6006



3.使用W4A16量化，调整KV Cache的占用比例为0.4，使用Python代码集成的方式运行internlm2-chat-1.8b模型。（优秀学员必做）

新建python文件pipeline\_kv1.py。

touch /root/pipeline\_kv1.py

打开pipeline\_kv.py，填入如下内容：

from lmdeploy import pipeline, TurbomindEngineConfig

# 调低 k/v cache内存占比调整为总显存的 20%

backend\_config = TurbomindEngineConfig(cache\_max\_entry\_count=0.4)

pipe = pipeline('/root/internlm2-chat-1\_8b',

backend\_config=backend\_config)

response = pipe(['Hi, pls intro yourself', '上海是'])

print(response)

保存后运行python代码：

python /root/pipeline\_kv.py

运行截图

